Journal of Biomedical Informatics 43 (2010) 435–441

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

TMA-TAB: A spreadsheet-based document for exchange of tissue microarray data based on the tissue microarray-object model

Young Soo Song [a], Hye Won Lee [b], Yu Rang Park [a], Do Kyoon Kim [a], Jaehyun Sim [a], Hyunseok Peter Kang [c], Ju Han Kim [a,d,*]

[a] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Republic of Korea
[b] Dept. of Molecular Genetics and Microbiology, College of Medicine, University of Florida, FL, USA
[c] Dept. of Pathology and Laboratory Medicine, Roswell Park Cancer Institute, Buffalo, NY, USA
[d] Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea

ARTICLE INFO

Article history:
Received 10 June 2009
Available online 14 October 2009

Keywords:
Database
Microarray data
Modeling
Tissue microarray

ABSTRACT

The importance of tissue microarrays (TMA) as clinical validation tools for cDNA microarray results is increasing, whereas researchers are still suffering from TMA data management issues. After we developed a comprehensive data model for TMA data storage, exchange and analysis, TMA-OM, we focused our attention on the development of a user-friendly exchange format with high expressivity in order to promote data communication of TMA results and TMA-OM supportive database applications. We developed TMA-TAB, a spreadsheet-based data format for TMA data submission to the TMA-OM supportive TMA database system. TMA-TAB was developed by simplifying, modifying and reorganizing classes, attributes and templates of TMA-OM into five entities: experiment, block, slide, core_in_block, and core_in_slide. Five tab-delimited formats (investigation design format, block description format, slide description format, core clinicohistopathological data format, and core result data format) were made, each representing the entities of experiment, block, slide, core_in_block, and core_in_slide. We implemented TMA-TAB import and export modules on Xperanto-TMA, a TMA-OM supportive database application, to facilitate data submission. Development and implementation of TMA-TAB and TMA-OM provide a strong infrastructure for powerful and user-friendly TMA data management.

1. Introduction

Tissue microarrays (TMA) are a promising array-based technology in cancer research and their importance in pathology is increasing due to their role in the clinical validation of cDNA microarrays [1]. TMA technology allows researchers to examine the expression of protein, DNA or RNA on hundreds or thousands of tissue samples while preserving morphology [2]. This increased throughput accelerates the discovery of important biologic markers compared to traditional marker studies using whole slide sections and has made this technology an essential tool in human protein profiling [3].

There is an enormous amount of data, including clinical and histopathological information associated with the cores in TMA blocks. This data grows exponentially even with a single experiment, which generates interpretation results for each core on a slide. Without powerful data management tools, the incredible volume of TMA data can be a burden to researchers, resulting in improper interpretation of data. For example, if data about the interpretation of the cores is recorded in one repository and data about the clinical and histopathological findings in another and there is no available informatics tool to integrate these data, one may try to do this manually, increasing chances of misinterpretation, especially without proper identifier and vocabulary management. Many TMA researchers typically work in laboratories without bioinformatics support and have difficulties managing TMA data.

In biomedical research, the development of standards, such as minimum information specification, data exchange format, and object model are essential to provide a solid basis for the development of data management applications. In the fields of cDNA microarray and proteomics, these efforts have been made by the Microarray Gene Expression Data (MGED) group and the Human Proteome Organization (HUPO), respectively (Table 1) [4–8]. These standards are successfully implemented and widely used, the typical examples being ArrayExpress in cDNA microarray and PEDRo in proteomics. Along with these trends, standards have also been proposed for TMAs.

* Corresponding author. Address: Division of Biomedical Informatics, Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea. Fax: +82 2 742 5947.
E-mail address: juhan@snu.ac.kr (J.H. Kim).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.10.001

Table 1
Comparison between development of data standards in biomedical research.

Data standards	cDNA microarray data	Proteomics data	TMA data
Minimum information specification	MIAME	MAIPE	TMA DES
Data model	MAGE-OM	PSI-OM	TMA-OM
XML format for data exchange	MAGE-ML	PSI-ML	TMA DES
Spreadsheet format for data exchange	MAGE-TAB	PRIDE proteomics harvest spreadsheet	Not available
Implementation	ArrayExpress	PEDRo	Xperanto-TMA

MIAME: minimal information about microarray experiment, MAGE-OM: microarray gene expression object model, MAGE-ML: microarray gene expression markup language, MAGE-TAB: microarray gene expression tabular, MAIPE: minimum information about a proteomics experiment, PSI-OM: proteomics standards initiative object model, PSI-ML: proteomics standards initiative markup language, PRIDE: proteomics identifications database, PEDRo: proteome experimental data repository, TMA DES: tissue microarray data exchange specification, TMA-OM: tissue microarray-object model.

The Association of Pathology Informatics proposed an open access TMA data exchange specification (TMA DES) as a format for sharing TMA data in 2003 [9]. TMA DES is a well-made XML document with a suitable structure that contains essential data elements of TMAs, such as experiment, block, slide and core in a hierarchical design and is very useful in the management of TMA data.

Our group proposed TMA-OM as a data model with integrity, flexibility and extensibility in dealing with TMA data [10]. TMA-OM provides a comprehensive model for storage, analysis and exchange of TMA data and also facilitates model-level integration with other biological models. During the development of TMA-OM, every kind of data and event that a TMA experiment can produce was thoroughly analyzed, including experiment design, block design, acquisition of clinical and histopathological data, block construction, slide cutting, staining, image acquisition, image analysis and management of the whole system. TMA-OM, having multidimensional features, can provide data necessary not only for researchers but also for technicians, block manufacturers, antibody producing companies and developers of TMA database systems. As the first application based on TMA-OM, a web-based database management system, Xperanto-TMA (available at http://xperanto.snubi.org/tma/) was implemented.

The TMA-OM supportive database has been suffered from the complexity of data models, long list of required elements, and low level of user-friendliness for the non-informatician pathologists. Instead of improving the user interface, we concluded that we needed a simpler, ease-to-understand representation of TMA data reflecting the perspective of a typical TMA researcher.

To overcome the limitations of TMA-OM, we designed a spreadsheet-based data exchange format for TMA data. There were three rationales for the development of a spreadsheet-based format. First, we tried to address the drawbacks of the TMA DES, which does not provide detailed instructions for clinical and histopathological data, with data structure of each document being dependent on the author, which creates the possibility that results of identical experiments might have different data structures. Moreover, because TMA DES is based on XML, it is not available to most researchers working in laboratories without bioinformatics support. Second, the multidimensional nature of TMA-OM is not suitable as a data exchange format and needs to be simplified for TMA data exchange. We created a new model for TMA data exchange by selecting and reorganizing the data elements in TMA-OM. The data exchange format based on this model should provide sufficient clinical and histopathological information to the level of granularity required for most TMA research. Third, spreadsheets are a useful data exchange format in biomedical research when experimental design is regular or simple. From our experience, most TMA research projects have a simple experimental design and a set of designs can be defined that encompass most projects. Moreover, spreadsheets are a very familiar format to most researchers and much TMA data is already stored in this format. This is not unique to TMAs. Spreadsheet-based data exchange formats including MAGE-TAB, PRIDE Proteomics Harvest Spreadsheet, and ISA-TAB, were developed for cDNA microarray, proteomics and combinations of omics-based experiments, respectively [11–13]. The spreadsheet format has also been used for partial uploading of TMA data in other TMA database systems [14,15]. The usefulness of a general format compared to a specific interface is that it gives more freedom to both researchers and developers without being limited to specific platforms.

In this article we propose TMA-TAB as a spreadsheet-based data exchange format for TMA data. TMA-TAB can be used for data collection, presentation, and communication between researchers or machines. It is easy-to-learn without any knowledge about bioinformatics. We also implemented an import and export interfaces to the TMA-OM supported web application, Xperanto-TMA. We expect that this will accelerate TMA workflow, promoting TMA research as a whole.

2. Methods

2.1. Conceptual schema

The first step in designing a simple and easy-to-learn format for data exchange was to determine the data elements of TMA experiments that are of concern to researchers. Most researchers are interested in how results of immunohistochemical assays correlate with the clinical and histopathological data annotations of each core section on a slide.

Next, we had to generalize those data elements into several representative entities. Experiment, block, slide, core_in_block and core_in_slide were chosen as five entities representing essential TMA data. Core_in_block and core_in_slide play a role in annotating clinical and histopathological data and interpreting results. Block and slide connect these two entities and experiment encompasses all of these entities. These five entities were partially implemented by the TMA DES although it did not divide core into core_in_block and core_in_slide [9]. Using these five entities, most of the concepts in TMA data important to researchers can be successfully described (Table 2). One of the advantages of introducing these entities is that these are very familiar concepts to researchers, hence enabling easy understanding of the structure and relationships of the entities.

We then generated attributes for each entity, which explain and describe the characteristics of each entity. Attributes were drawn from the classes, attributes and templates of TMA-OM, and these were reorganized, simplified and modified based on the needs of researchers. This process occurred in four steps. First, only classes containing real TMA data were selected while classes representing processes or events were excluded.

Second, the remaining classes were clustered into five entities and related classes were combined to produce new attributes if this process did not cause severe information loss. For example, the TMA-OM's TumorInfo class in the HisoPathol package having

Table 2
Overall features of TMA-TAB and its relationship with TMA-OM.

Entities in TMA-TAB	Data format in TMA-TAB	Data contents	Packages in TMA-OM (percentage of classes represented by TMA-TAB among total classes of each package)
Experiment	IDF	Title, ExpType, ExpFactor, Description, ExternalLink	Experiment (100%)
Block	BDF	BlockIdentifier, NumOfRow, NumOfCol, CoreSize, BlockConstructionProtocol, BlockCreationDate, Description, ExternalLink	Block (25%), BlockDesign (100%)
Slide	SDF	SlideIdentifier, SlideStain, SlideTestCategory, SlideSerialNumber, SlideProtocol, SlideCutDate, SlideStainDate, BlockIdentifier, Description, ExternalLink	Array (67%), BioAssay (9%), Reporter (25%)
Core_in_block	CCDF	43 templates dependent on tissue and cancer types	DesignElement (33%), BioMaterial (43%), HistoPathol (100%), ClinInfo (56%)
Core_in_slide	CRDF	Availability, PercentOfTissueStaining, TissueIntensity, NumberOfNucleiCounted, EvaluationCategory, StainingCompartment, StainingPattern, CoreType, InterpretationProtocol, Description, SlideIdentifier, PosRow and PosCol	BioAssayData (6%), QuantitationType (71%)
Absent	Protocol format	ProName, ProType, Description	Protocol (10%)

IDF: investigation description format, BDF: block description format, SDF: slide description format, CCDF: core clinicohistopathologic data format, CRDF: core result data format.

classes, Tstage, Nstage, Mstage, BasicHistoPathol, NstageInfo, TstageInfo, MstageInfo, TNMstage, pathologist_reviewed and tumorStageCodeType, can be modified and simplified as attributes of Tstage, Mstage, Nstage and pathologists in the core_in_block entity through unification of associated classes. Every class of the TMA-OM was investigated in this way.

Third, each attribute was evaluated as to whether the data it represented was really practical in the TMA experiment. As a result of this process, 53% of classes and 64% of attributes in the TMA-OM are represented by TMA-TAB. Excluded classes represent an events or a processes and excluded attributes describe technical details, most likely beyond the interest of researchers.

Fourth, 43 premade templates in TMA-OM for describing organ-specific specimen information were restructured into sets of categories and values and the categories were added to the attributes of core_in_block. For example, a template in TMA-OM for gastrointestinal lymphoma consists of three common data element (CDE) groups (Macroscopic, Microscopic, Histologic), 12 categories under the CDE groups (HistologicType_NonHodgkinLymphoma, HistologicType_B-cellLymphoma, HistologicType_T-cellLymphoma, etc.), and 75 values under the categories (B-cellLymphoma, T-cellLymphoma, Hairy cell leukemia, etc.). The template is restructured by removing the CDEs, CDE groups, and the hierarchical structures of the categories and subcategories. The categories, HistologicType_B-cellLymphoma and HistologicType_T-cellLymphoma, for example, are subcategories of HistologicType_NonHodgkinLymphoma. Because hierarchical information is hard to apply to TMA-TAB and the permissible values for HistologicType_B-cellLymphoma and HistologicType_T-cellLymphoma are mutually exclusive with each other and exhaustive to the super-category, HistologicType_NonHodgkinLymphoma, these two subcategories can be unified and merged into HistologicType_NonHodgkinLymphoma without information loss. Each step involves no information loss because the permissible values in the pathologic diagnosis of a sample for HistologicType_B-cellLymphoma and HistologicType_T-cellLymphoma are mutually exclusive and exhaustive to HistologicType_NonHodgkinLymphoma. Then each restructured category was entered as an attribute into the entity of core_in_block. The values of each category are used for determining the permissible values of each cell (see Section 2.3).

After generation of attributes, we defined rules of relationship between the entities, listed below.

1. An instance of a block is owned by one or more instances of experiments.
2. An instance of a slide originates from an instance of a block.
3. An instance of a core_in_block is owned by an instance of a block.
4. An instance of a core_in_slide originates from an instance of a core_in_block and also owned by an instance of a slide.

If two entities are related, each entity should have attributes both for identifying self and for referring to the other entity that it owns or originates from. In this way, entities can refer to each other. Referring data from an instance of core_in_slide to an instance of core_in_block is a reflection of a real world event of TMA data processing where researchers analyzing a core in a TMA slide find the corresponding clinical and histopathologic data annotated to a core with the same coordinates in the source block.

2.2. Formalization of TMA-TAB from conceptual schema

We created five tab-delimited files from the premade conceptual schema that preserved their structure. These are investigation description format (IDF) from the experiment in the conceptual schema, block description format (BDF) from the block, slide description format (SDF) from the slide, core clinicohistopathologic data format (CCDF) from the core_in_block and core result data format (CRDF) from the core_in_slide (Table 2).

Headers in the first row correspond to the attributes in the conceptual schema. TMA data is inserted into the cells under the headers. Each row of data corresponds to one instance of an entity.

In the case of CCDF, it was not reasonable to use all the attributes taken from the conceptual schema because important clinical and histopathologic data vary depending on the tissue examined and the type of cancer. We created, therefore, 43 types of CCDF templates for 43 cancers according to the College of American Pathologists (CAP) Cancer Protocols and checklists so that researchers can select a template best describing the experiment.

Besides these five formats, attributes describing protocols or procedures in conceptual schema were organized separately into protocol formats. These are block construction protocol, slide protocol, pretreatment protocol for antibody or probe, fixation protocol, surgical procedures, and slide reading protocol. Though the same information can be provided regardless of whether data about protocols or procedures are stored independently (protocol formats) or in association with IDF, BDF, SDF, CCDF or CRDF, this reduces the potential redundancy of TMA-TAB.

2.3. Ontology

Vocabularies used in TMA-TAB are taken from MGED Ontology, TMA DES, terms from MISFISHIE, CDEs of CAP Cancer Protocols and NCI CDEs [9,16,17] as in TMA-OM [10]. Permissible values of each cell were determined by the header and are specified in the docu-

ment of specifications on TMA-TAB [18]. In brief, the values were selected to be made both convenient to use and compatible with that of implemented TMA-OM. If the header corresponded to a category of a template in TMA-OM, the values under the category in the template were used for the permissible values of the cell under the header, slightly modified for convenience if necessary.

2.4. Application

Finally we implemented TMA-TAB on Xperanto-TMA, a web-based TMA database application using TMA-OM, allowing researchers to submit TMA data by simply uploading TMA-TAB files.

3. Results

3.1. Structure of TMA-TAB

TMA-TAB consists of five tab-delimited files (IDF, BDF, SDF, CCDF and CRDF) and additional protocol files (Table 2). According to definitions from the RSBI working group, 'investigation' is a self-contained unit of scientific inquiry with a holistic hypothesis or objective and 'assay' is a part using particular technologies [19]. TMA-TAB can contain data on only one investigation, but more than one assay can be included under one investigation.

Each file in TMA-TAB has headers in the first row and TMA data can be inserted starting from the second row (Fig. 1). For the submission of TMA-TAB into a TMA database, the relationship with preexisting data should be considered. For example in Xperanto-TMA, if the value of the 'Title' column in IDF is 'MTA-1 expression in colon cancer' and another experiment with the same title has been already registered in the database, users are prevented from submitting the TMA data under the same title. Users should check if the data to be submitted is already stored in the TMA system. If the files represent a different experiment the Title attribute should be changed. With this policy, each experiment in TMA database system has a unique title, preserving data integrity.

The following is a brief description of each format. For more detailed information and examples of TMA-TAB, please refer to the document of specifications (Suppl_TMA_TAB_Specification.htm, Suppl_example_colorectal.xls and Suppl_UML.htm, available at http://xperanto.snubi.org/TMA/suppl/).

3.1.1. Investigation description format (IDF)

IDF describes the overall outline of an experiment including experimental factor, design and type. Because TMA-TAB can include only one instance of a TMA experiment, IDF has only two

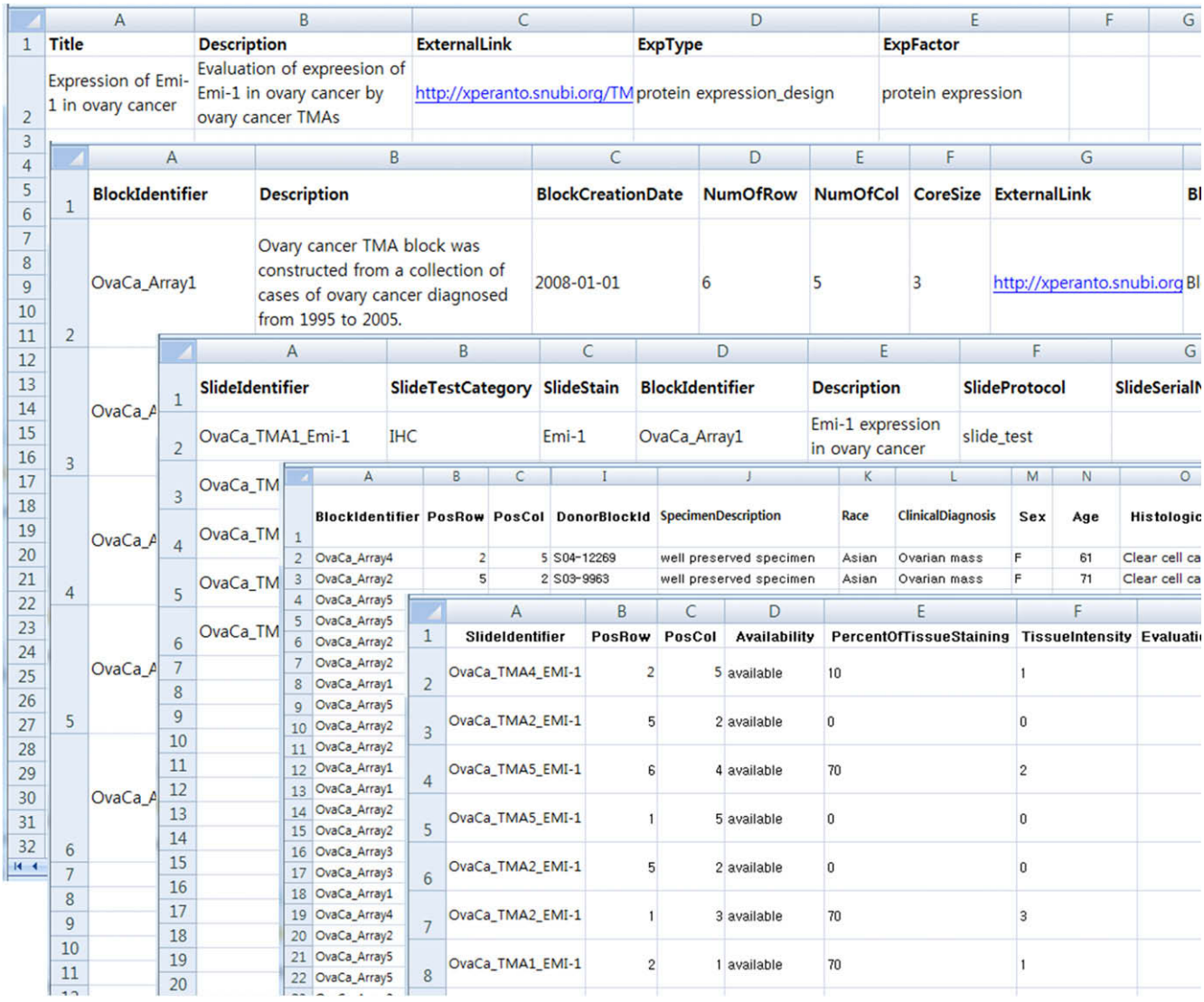

Fig. 1. Example of TMA-TAB usage with an ovarian cancer template.

rows, headers in the first row and data in the second row. The headers are Title, ExpType, ExpFactor, Description, and ExternalLink. No additional headers are permitted. Controlled vocabularies and ontologies including the MGED Ontology, TMA DES, terms from MISFISHIE, CDEs of CAP Cancer Protocols and NCI CDEs are applied for the values of ExpType and ExpFactor. Any string can be applied to describe the Title, Description and ExternalLink, except that the values of Title should be unique among the experiments stored in a system for the purpose of eliminating conflicts. All permissible values for each cell in the TMA-TAB format are described in the specification file (http://xperanto.snubi.org/tma/Suppl/Suppl_TMA_TAB_Specification.htm).

3.1.2. Block description format (BDF)

BDF contains overall information about blocks such as name, numbers of rows and columns, and core size. For the submission of TMA-TAB to TMA database system, the value of BlockIdentifier should be unique: and data with a block identifier that exists in the database cannot be resubmitted. Unlike TMA-OM, the unit of CoreSize is already determined as mm.

3.1.3. Slide description format (SDF)

SDF describes the general information of each slide, such as slide name, stain and slide test category. For the submission of TMA-TAB, SlideIdentifier should be unique under a single experiment. This means if slides belong to different experiments, the same SlideIdentifier is allowed. The value of SlideStain is the name of the antibody, probe or lectin. For submission, the staining material should be registered first, providing information about the target molecule, type of staining, staining compartment and reporter provider. BlockIdentifier of SDF refers to the name of the block the slide originates from: information on the block may already exist in the database or be submitted at the same time.

3.1.4. Core clinicohistopathologic data format (CCDF)

CCDF contains information on tissue cores and annotated clinical and histopathological information. Unlike other formats, CCDF has 43 templates depending on the tissue and type of cancer and user-defined data elements can be added to any existing templates. For example, a template for colorectal cancer has 38 headers, including BlockIdentifier, PosRow, PosCol, SpecimenId, Fixation, FixationProtocol, Sex, Age, Histology, HistologicGrade and TumorSize. Although there is no single column for the identifier of the core, the combination of BlockIdentifier, PosRow (position of row), and PosCol (position of column) fulfills this role. The unit of TumorSize is designated as cm.

When describing microscopic configuration of a tumor, 'infiltrating' and 'invasive' can refer to similar characteristics, but in TMA-TAB, 'infiltrating' is a permissible value while 'invasive' is not permissible in the MicroscopicConfiguration data element. This allows clear description of TMA data that both humans and machines can understand. Table 3 shows an example of a CCDF of colorectal cancer (Due to the limitation of space, only part of the CCDF is shown. An example with the entire CCDF is provided in the Supplementary material) [18].

3.1.5. Core result data format (CRDF)

CRDF contains experimental data on the cores of TMA slides. Headers include Availability, PercentOfTissueStaining, TissueIntensity, NumberOfNucleiCounted, EvaluationCategory, StainingCompartment, StainingPattern, CoreType, InterpretationProtocol, Description, SlideIdentifier, PoswRow and PosCol. The combination of SlideIdentifier, PosRow, and PosCol serve as a unique identifier.

3.2. Implementation of TMA-OM and TMA-TAB as Xperanto-TMA

Xperanto-TMA is a web-based application using MySQL 4.1 and based on TMA-OM [20]. The relational schema is derived from TMA-OM by object-relational mapping. The experiment-friendly interface of Xperanto-TMA was designed with the general workflow of a TMA experiment in mind. Xperanto-TMA accommodates a controlled vocabulary and a template-driven data management system providing design and registry functionalities. Since Xperanto-TMA was implemented in 2006, several functions have been added to the initial system: the complete list of features is described below.

3.2.1. Data submission

The data submission function aims to provide an accurate recording tool by adopting aspects of structured data entry such as controlled vocabularies and pre-defined data elements. Xperanto-TMA provides two ways of data submission: (1) editing online submission forms for experiment, slide, block and core data and (2) uploading TMA-TAB files.

When submitting data by editing online submission forms, users should enter information about experiment, block and slide before the submission of data on core_in_block or core_in_slide. Users can insert TMA data either by typing or by choosing one of items from the selection box to use the controlled vocabulary.

If users submit data by uploading TMA-TAB, they can select from five scenarios. These scenarios are developed for the user's convenience for situations when the whole set of experiments is not completed and only part of the data is available but users want to upload the data that they have. For example, when TMA blocks and annotated clinical and histopathologic data have been prepared but the slides have not been stained yet, users can upload BDF and CCDF by selecting the fifth scenario. After uploading, the system automatically validates formats, data relevance, and relationship between each format, preventing incorrect or discrepant

Table 3
An example of a part of a CCDF for colorectal cancer.

BlockIdentifier	PosRow	PosCol	SpecimenId	Sex	Age	DiagnosisDate	OperationName	Histology	TumorSite	TumorSize
Colon_Array73	1	1	S1991–155	F	41	1991–01-09	Abdominoperineal resection	Adenocarcinoma|Mucinous[a]	Rectum	8
Colon_Arrray73	1	2	S1991–325	M	72	1991–01-16	Abdominoperineal resection	Adenocarcinoma	Rectum	10
Colon_Arrray73	1	3	S1991–808	M	53	1991–02-06	Abdominoperineal resection	Adenocarcinoma	Rectum	4
Colon_Arrray73	2	1	S1991–1000	M	51	1991–02-09	Sigmoidectomy	Adenocarcinoma	Sigmoid colon	3.5
Colon_Arrray73	2	2	S1991–1131	M	67	1991–02-19	Sigmoidectomy	Adenocarcinoma	Sigmoid colon	3
Colon_Arrray73	2	3	S1991–9999	M	58	1991–02-20	Abdominoperineal resection	Adenocarcinoma	Rectum	5
Colon_Arrray74	1	1	S1991–559	F	41	1991–02-22	Abdominoperineal resection	Adenocarcinoma	Rectum	9
Colon_Arrray74	1	2	S1991–1225	M	72	1991–03-03	Left hemicolectomy	Adenocarcinoma	Descending colon	9
Colon_Arrray74	2	1	S1991–2808	M	53	1991–03-03	Abdominoperineal resection	Adenocarcinoma	Rectum	10
Colon_Arrray74	2	2	S1991–1011	M	51	1991–03-05	Left hemicolectomy	Adenocarcinoma	Descending colon	2

[a] If multiple values are allowed in a cell, use '|' as a delimiter. To find the fields where multiple values, refer to the document of specification in the Supplementary material (http://xperanto.snubi.org/TMA/suppl/Suppl_TMA_TAB_Specification.htm).

values from being submitted to the system. During the submission of TMA-TAB, one can also describe user-defined terms.

3.2.2. Data export: text and XML

Users can export the data for each experiment as well as tissue information into tab-delimited text and XML files conforming to the TMA DES. The exported file contains all information about the experiment including array, clinical and histopathological information.

3.2.3. Controlled vocabulary

Xperanto-TMA utilizes controlled vocabularies including MGED Ontology [17], 80 tags of TMA DES, terms from MISFISHIE for TMA experiment procedures, and CDEs extracted from CAP Cancer Protocols and NCI CDEs for clinical and histopathologic information. CDEs for clinical and histopathologic information are under the control of a system administrator but user-defined CDEs can be added with sysadmin approval. Allowing user-defined CDEs may eventually require a central 'standard' repository of CDEs that are widely accepted by the TMA data management community. In the mean time, collaborators can run regional CDE repositories with administrative control and communicate periodically.

3.2.4. Template management

The template is a form composed of common data elements (CDEs) which are metadata to describe data. Researchers can use pre-defined templates provided in Xperanto-TMA but also submit new templates by organizing CDEs. Xperanto-TMA provides 43 templates based on CAP cancer protocol to guide entries for each cancer type, and accommodates an ISO 11179-supported data registry for detailed description of data elements. These functions for templates and controlled vocabularies allow users to systematically submit and manage data.

4. Discussion

As mentioned above, TMA-TAB was developed and implemented with tight relationships with TMA-OM to provide a strong infrastructure for powerful and user-friendly TMA data management. We addressed several issues related to these objectives.

4.1. Design and implementation of TMA-TAB

We selected spreadsheets as a basic format of TMA-TAB because the table structure of spreadsheets is suitable to TMA data and easy to manage. Though the spreadsheet is a useful format for researchers, if there is no way to validate the data, this can become a serious problem [21]. Xperanto-TMA provides data validation on submission of TMA-TAB. Data validation is achieved in three ways: format validation, cell value validation and validation of logical relationships among formats. In format validation, the application automatically checks for inconsistencies with IDF, BDF, SDF, CCDF, CRDF or protocol formats. Each cell is checked for permissible values in cell value validation. This allows the TMA database system to adhere to controlled vocabularies. When validating the logical relationship between formats, for example, if a row in CCDF has a BlockIdentifier that does not exist, the relationship between CCDF and BDF is logically wrong. This kind of error can happen when users insert the wrong identifier value, in which case Xperanto-TMA can detect and notify the user of the logically incorrect relationship.

We did not provide detailed information about slide-staining material, which seems to be a rather technical issue, to focus our attention on the essential entities such as experiment, block, slide, core_in_block and core_in_slide, which are more directly related to the evaluation of experiment. When implementing the interface for TMA-TAB, this issue can be dealt in various ways according to user decisions. In Xperanto-TMA, when submitting experiments utilizing new staining materials, users should complete the forms about staining materials prior to submitting the TMA-TAB.

4.2. Implications of TMA-TAB and TMA-OM

The use of TMA-TAB is not limited to TMA data submission to a database system. Because TMA-TAB includes almost all the information necessary for the evaluation of an experiment, researchers can use it to easily understand the entire process of an experiment and evaluate the data obtained, making it a useful tool for communication between researchers. It also guides the process of data collection because it informs researchers about generally required clinical and/or histopathological information and what should be included when reporting results.

TMA-TAB was developed by an extensive analysis and restructuring of TMA-OM from the perspective of TMA researchers. Data representation in TMA-TAB may appear less explicit than that of TMA-OM, however, this can be transformed into a clearly explicit representation. Such transformation is possible because the nature of TMA technology and knowledge about clinical medicine and pathology can provide implicit links between data elements in TMA-TAB. For example, when the value of a cell in HistologicType section of TMA-TAB is "Diffuse large B-cell lymphoma", the value should be mapped into category, "B-cell lymphoma", even though it is not explicitly stated in the TMA-TAB that "Diffuse B-cell lymphoma" is a part of "B-cell lymphoma". This mapping is supported by TMA-OM, which provides knowledge about the classification of lymphoma. To present every data element completely explicitly is almost impractical from a human perspective, especially for domain experts, but this step is necessary in data storage. By implementing both TMA-OM and TMA-TAB, we preserve explicit data semantics while making the process of data exchange more convenient for researchers.

4.3. Features distinguishing Xperanto-TMA from the other public TMA database systems

There are several major features that distinguish Xperanto-TMA from the other public TMA database systems recently reported in the literatures [1,14,15,22,23]. First, Xperanto-TMA can explicitly represent data elements in the system because it is based on TMA-OM, an object model which was designed after thorough investigations of TMA data. It also prevents logically incorrect data entry. Explicit representation of TMA data also helps development of other related applications. Second, TMA-OM was developed through the seamless integration of existing standards in closely related domains. This means that TMA-OM can grow together with other standards as the standards are upgraded. Third, the functions of Xperanto-TMA can be easily extended with other database systems for "-omics" technologies. Fourth, through the implementation of TMA-TAB, Xperanto-TMA can upload data of a TMA experiment data as a spreadsheet format. Though spreadsheets have been used in Cruella [14] and Stanford Tissue Microarray Database [15] to describe results of an experiment, TMA-TAB contains the full range of information of an experiment including block, slide, clinical and histopathological information as well as results of interpretation.

4.4. Evaluation study for TMA-TAB

We performed a very brief survey to evaluate TMA-TAB with five pathologists, who had experience with TMA experiments but did not have any bioinformatics knowledge. All of them used MS

Excel spreadsheets for the storage and management of TMA data mainly because they had no concept of an informatics applications based on database management system. We asked the pathologists to fill in two of each of the MS Excel documents in TMA-TAB format from two TMA datasets. Given basic instructions on TMA-TAB and the specification with an example TMA-TAB file, all of them had no difficulty in producing documents in TMA-TAB format with very high percentage of correctness. After using TMA-TAB, all of them completed a questionnaire (see supplement) about its practicality and user-friendliness. Most of them said that the structure of TMA-TAB appeared to be optimal for the description of TMA data, time to complete a document describing a TMA experiment would be shortened, and they would readily use TMA-TAB if a TMA-TAB 'editor' were provided. Four of them answered that TMA-TAB is user-friendly, but one answered that it was not user-friendly without an editor.

4.5. Necessities for the management of user-defined terms in TMA-TAB and Xperanto-TMA

Although the CAP cancer protocols provide a standard format for reporting cancer specimens in the practice of surgical pathology, they do not always reflect the interests of researchers in clinical and histopathologic findings. For example, no structured format is provided in current CAP cancer protocols to describe radiologic findings of brain tumor. Many studies of brain tumor have investigated the correlation between histopathological and radiologic findings. To describe such data elements not defined in CAP cancer protocols, user-defined data elements are allowed in TMA-TAB, initially limited to the user group, he or she belongs to. If used frequently, we will update vocabulary lists to specifically describe these data elements and the templates of TMA-TAB will be updated accordingly.

4.6. Future works on TMA-TAB

We expect the user-friendliness of TMA-TAB would be enhanced by providing a TMA-TAB specific editor. This would be designed to easily produce a TMA-TAB file conforming to all the requirements described in the specification of TMA-TAB. Using this editor would shorten the time required to produce a TMA-TAB file.

We also have plans for extracting pathology report data in TMA-TAB format from the electrical medical record. This would require extensive text mining techniques due to the fact that many pathologist still provide reports in an unstructured format.

5. Conclusions

We presented both machine-oriented and human-oriented TMA data representations as TMA-OM and TMA-TAB, respectively, and demonstrated that these can be used interchangeably. TMA-TAB may facilitate the process of data submission to TMA-OM supportive database systems. Xperanto-TMA, a web-based application on which TMA-OM and TMA-TAB were completely implemented, is a robust, easily adaptable database system with strong ability of data management. These achievements were possible because the TMA-TAB is based on a thorough analysis of TMA workflow. The development of TMA-TAB, which is essentially a simplified transformation of TMA-OM, was rather easier because all of the up-front work has been completed. This example shows the benefits of a comprehensive data model in the development of other models or applications. We will continue to develop easily adaptable applications based on the strong infrastructure of TMA-OM and TMA-TAB.

Acknowledgments

This study was supported by grant from the Korea Health 21 R&D Project, Ministry of Health, Welfare and Family Affairs, Republic of Korea (A040163). Educational training of Y.S.S. was supported by grant from the Korea Health 21 R&D Project, Ministry of Health, Welfare and Family Affairs, Republic of Korea (A030001).

References

[1] Viti F, Merelli I, Caprera A, Lazzari B, Stella A, Milanesi L. Ontology-based, tissue microarray oriented, image centered tissue bank. BMC Bioinformatics 2008;9(Suppl. 4):S4.
[2] Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 1998;4:844–7.
[3] Chung JY, Braunschweig T, Tuttle K, Hewitt SM. Tissue microarrays as a platform for proteomic investigation. J Mol Histol 2007;38:123–8.
[4] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365–71.
[5] Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, et al. PEDRo: a database for storing, searching and disseminating experimental proteomics data. BMC Genomics 2004;5:68.
[6] Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, et al. A design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 2002;3. RESEARCH0046.
[7] Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nat Biotechnol 2003;21:247–54.
[8] Taylor CF, Paton NW, Lilley KS, Binz PA, Julian Jr RK, Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 2007;25:887–93.
[9] Berman JJ, Edgerton ME, Friedman BA. The tissue microarray data exchange specification: a community-based, open source tool for sharing tissue microarray data. BMC Med Inform Decis Mak 2003;3:5.
[10] Lee HW, Park YR, Sim J, Park RW, Kim WH, Kim JH. The tissue microarray object model: a data model for storage, analysis, and exchange of tissue microarray experimental data. Arch Pathol Lab Med 2006;130:1004–13.
[11] Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 2006;7:489.
[12] Jones P, Cote RG, Cho SY, Klie S, Martens L, Quinn AF, et al. PRIDE: new developments and new datasets. Nucleic Acids Res 2008;36:D878–83.
[13] Sansone SA, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, et al. The first RSBI (ISA-TAB) workshop: can a simple format work for complex studies? OMICS 2008;12:143–9.
[14] Cowan JD, Rimm DL, Tuck DP. Cruella: developing a scalable tissue microarray data management system. Arch Pathol Lab Med 2006;130:817–22.
[15] Marinelli RJ, Montgomery K, Liu CL, Shah NH, Prapong W, Nitzberg M, et al. The stanford tissue microarray database. Nucleic Acids Res 2008;36:D871–7.
[16] Deutsch EW, Ball CA, Berman JJ, Bova GS, Brazma A, Bumgarner RE, et al. Minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE). Nat Biotechnol 2008;26:305–12.
[17] Stoeckert CJ, Parkinson H. The MGED ontology: a framework for describing functional genomics experiments. Comp Funct Genomics 2003;4:127–32.
[18] Web supplementary information. Available at: http://xperanto.snubi.org/TMA/suppl.
[19] Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, et al. The functional genomics experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol 2007;25:1127–33.
[20] Xperanto-TMA. Available at: http://xperanto.snubi.org/TMA/.
[21] Jameson D, Garwood K, Garwood C, Booth T, Alper P, Oliver SG, et al. Data capture in bioinformatics: requirements and experiences with Pedro. BMC Bioinformatics 2008;9:183.
[22] Sharma-Oates A, Quirke P, Westhead DR. TmaDB: a repository for tissue microarray data. BMC Bioinformatics 2005;6:218.
[23] Thallinger GG, Baumgartner K, Pirklbauer M, Uray M, Pauritsch E, Mehes G, et al. TAMEE: data management and analysis for tissue microarrays. BMC Bioinformatics 2007;8:81.